Analytics and Big Data: Terms and Tools

H Frank Cervone, PhD, CPHIMS
School of Public Health, University of Illinois at Chicago
School of Information, San Jose State University
DO WE HAVE ANY ACTIONABLE ANALYTICS FROM OUR BIG DATA IN THE CLOUD?

YES, THE DATA SHOWS THAT MY PRODUCTIVITY PLUNGES WHENEVER YOU LEARN NEW JARGON.

MAYBE IN-MEMORY COMPUTING WILL ACCELERATE YOUR APPLICATIONS.

PLUNGE, PLUNGE, PLUNGE.
Defining big data

A greater scope of information: 18%
New kinds of data and analysis: 16%
Real-time information: 15%
Data influx from new technologies: 13%
Non-traditional forms of media: 13%
Large volumes of data: 10%
The latest buzzword: 8%
Social media data: 7%

Respondents were asked to choose up to two descriptions about how their organizations view big data from the choices above. Choices have been abbreviated, and selections have been normalized to equal 100%. Total respondents=1144.
Neither snake oil nor silver bullet
Hidden patterns
Hidden correlations
Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »
(1) Collect real-time data.  
(2) Process data as it flows.  
(3) Explore and visualize.

Multiple Data Feeds

Calculate  Transform  Process  Augment

Dashboards, Reports

Ad-hoc Queries

Image courtesy of PracticalAnalytics
Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

Hbase
Columnar Store

Hive
SQL Query

R Connectors
Statistics

Mahout
Machine Learning

Pig
Scripting

Oozie
Workflow

Squoop
Data Exchange

Zookeeper
Coordination

Flume
Log Collector
Hadoop Distributed File System
Zookeeper
Yarn Map Reduce

• Map
  • Split input
  • Process local data and create key-value pairs

• Shuffle
  • Transfer results from mappers to reducers
  • Merge results from each partition by key

• Reduce
  • Read all results for each key
  • Perform application logic on data
Flume

Agent

Data

Source

Consume

Channel

Hold

Sink

Deliver

ExternalRepository
Pig Latin

A = LOAD 'file1' AS (x, y, z);
B = LOAD 'file2' AS (t, u, v);
C = FILTER A by y > 0;
D = JOIN C BY x, B BY u;
E = GROUP D BY z;
F = FOREACH E GENERATE
group, COUNT(D);
STORE F INTO 'output';

Logical Plan

LOAD (x, y, z)
FILTER (x, y, z)
JOIN (x, y, z, t, u, v)
GROUP (group, ((x, y, z, t, u, v)))
FOREACH (group, count)
STORE (group, count)
Mahout
Provisioning, managing, and monitoring clusters
<table>
<thead>
<tr>
<th>CRS_SL</th>
<th>CRS_NI</th>
<th>CRS_CRN</th>
<th>CRS_TITLE</th>
<th>PREREQUISITES</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSTT</td>
<td>400</td>
<td>BSTT 400</td>
<td>Biostatistics I</td>
<td></td>
</tr>
<tr>
<td>BSTT</td>
<td>401</td>
<td>BSTT 401</td>
<td>Biostatistics II</td>
<td>BSTT 400</td>
</tr>
<tr>
<td>BSTT</td>
<td>494</td>
<td>BSTT 494</td>
<td>Intro to Data Analysis w/R</td>
<td></td>
</tr>
<tr>
<td>BSTT</td>
<td>505</td>
<td>BSTT 505</td>
<td>Logist Regr &amp; Survivl Analysis</td>
<td>BSTT 400</td>
</tr>
<tr>
<td>BSTT</td>
<td>505</td>
<td>BSTT 505</td>
<td>Logist Regr &amp; Survivl Analysis</td>
<td>BSTT 401</td>
</tr>
<tr>
<td>BSTT</td>
<td>506</td>
<td>BSTT 506</td>
<td>Design of Clinical Trials</td>
<td>BSTT 400</td>
</tr>
<tr>
<td>BSTT</td>
<td>506</td>
<td>BSTT 506</td>
<td>Design of Clinical Trials</td>
<td>BSTT 401</td>
</tr>
<tr>
<td>BSTT</td>
<td>507</td>
<td>BSTT 507</td>
<td>Sampling/Est Methods in PH</td>
<td>BSTT 401</td>
</tr>
<tr>
<td>BSTT</td>
<td>507</td>
<td>BSTT 507</td>
<td>Sampling/Est Methods in PH</td>
<td>BSTT 502</td>
</tr>
<tr>
<td>BSTT</td>
<td>523</td>
<td>BSTT 523</td>
<td>Biostatistics Methods I</td>
<td>BSTT 524</td>
</tr>
<tr>
<td>BSTT</td>
<td>524</td>
<td>BSTT 524</td>
<td>Biostatistics Laboratory</td>
<td>BSTT 523</td>
</tr>
<tr>
<td>BSTT</td>
<td>525</td>
<td>BSTT 525</td>
<td>Biostatistics Methods II</td>
<td>BSTT 523</td>
</tr>
<tr>
<td>BSTT</td>
<td>525</td>
<td>BSTT 525</td>
<td>Biostatistics Methods II</td>
<td>BSTT 524</td>
</tr>
<tr>
<td>BSTT</td>
<td>535</td>
<td>BSTT 535</td>
<td>Categorical Data Analysis</td>
<td>BSTT 525</td>
</tr>
<tr>
<td>BSTT</td>
<td>535</td>
<td>BSTT 535</td>
<td>Categorical Data Analysis</td>
<td>STAT 411</td>
</tr>
<tr>
<td>BSTT</td>
<td>536</td>
<td>BSTT 536</td>
<td>Survival Analysis</td>
<td>BSTT 525</td>
</tr>
<tr>
<td>BSTT</td>
<td>536</td>
<td>BSTT 536</td>
<td>Survival Analysis</td>
<td>STAT 411</td>
</tr>
<tr>
<td>BSTT</td>
<td>537</td>
<td>BSTT 537</td>
<td>Longitudinal Data Analysis</td>
<td>BSTT 525</td>
</tr>
<tr>
<td>BSTT</td>
<td>537</td>
<td>BSTT 537</td>
<td>Longitudinal Data Analysis</td>
<td>STAT 411</td>
</tr>
<tr>
<td>BSTT</td>
<td>538</td>
<td>BSTT 538</td>
<td>Biostatistical Consulting</td>
<td>BSTT 525</td>
</tr>
<tr>
<td>BSTT</td>
<td>550</td>
<td>BSTT 550</td>
<td>Biostatistical Investigations</td>
<td>BSTT 535</td>
</tr>
<tr>
<td>BSTT</td>
<td>550</td>
<td>BSTT 550</td>
<td>Biostatistical Investigations</td>
<td>BSTT 536</td>
</tr>
<tr>
<td>BSTT</td>
<td>550</td>
<td>BSTT 550</td>
<td>Biostatistical Investigations</td>
<td>BSTT 537</td>
</tr>
<tr>
<td>BSTT</td>
<td>550</td>
<td>BSTT 550</td>
<td>Biostatistical Investigations</td>
<td>BSTT 538</td>
</tr>
</tbody>
</table>
Heart Failure Readmissions

30-Day Readmits: 20%
- Target: 20%
- 465 Discharges (Nov 21 to Dec 21)
- Readmits: 49

90-Day Readmits: 30%
- Target: 27%
- 1200 Discharges (Jul 24 to Oct 22)
- Readmits: 254

90-Day ER Utilization: 15%
- Target: 22%
- 1200 Discharges (Oct 24 to Oct 22)
- Visits: 177
- Stays: 63

90-Day Observation Stay: 10%
- Target: 22%
- 1200 Discharges (Oct 24 to Oct 22)

Interventions:
- Medication Reconciliation: 56%
- Follow Up Phone Call: 32%
- Discharge Appointment: 15%
- All Interventions: 2%

Readmissions Over Time:
- Graph showing trends in readmissions from 2007 to 2014.
Thank you

H Frank Cervone
fcervone@uic.edu